

# Metagenome Analysis With MG-RAST

Folker Meyer, PhD

Argonne National Laboratory  
and  
University of Chicago

Palm Springs, March 2013

<http://metagenomics.anl.gov>

# Acknowledgements

## Team:

- Dion Antonopoulos
- Daniela Bartels
- Jared Bischof
- Dan Braithewaite
- Sarah O'Brien
- Adina Chuang-Howe
- Narayan Desai
- Mark Domanus
- Mark d' Souza
- Katya Drybinski
- Elizabeth M. Glass
- Wolfgang Gerlach
- Kim M. Handley
- Travis Harrison
- Kevin Keegan
- Tobias Paczian
- Hunter Matthews
- Sarah Owens
- Wei Tang
- Will Trimble
- Andreas Wilke
- Jared Wilkening

## Major Collaborators:

- A. Arkin (Berkeley)
- E. Chang (UChicago)
- Dawn Field (Oxford)
- F.-O. Glöckner (MPI Bremen)
- Jack Gilbert (Argonne)
- Jeff Grethe (CalIT2, CAMERA)
- Sarah Hunter / Guy Cochrane (EBI)
- Ken Kemner (Argonne)
- Rob Knight (Colorado)
- Nikos Kyrpides (DOE JGI)
- J. Tiedje (MSU)
- Owen White (UMaryland, HMP DACC)



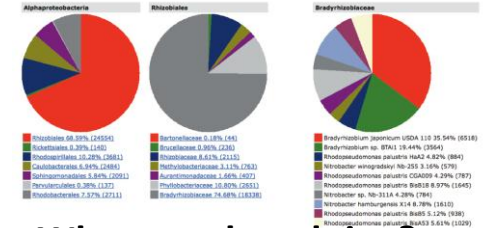
# Microbial community science

Who are they?

## → Discovery of novel functions

Environmental clone libraries (“functional metagenomics”)

- Sanger sequencing of BAC clones with env. DNA
- low throughput but supports in vitro screens



## → Ecology

Amplicon studies (single gene studies, 16s rDNA)

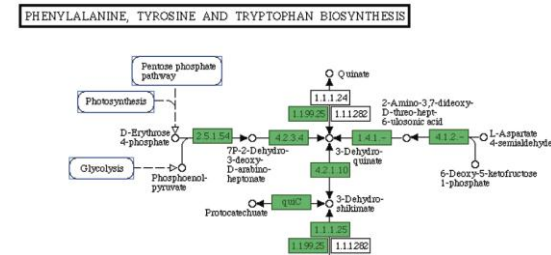
- Sequencing of PCR amplified ribosomal genes
- sequence quality limited (→ rare biosphere debate)
- often can't distinguish between individual strains

What are they doing?

## → Ecology and Discovery

Shotgun metagenomics

- “random shotgun DNA sequencing applied directly to environmental samples”



# What is MG-RAST

- Web based
  - Upload, process, share and publish microbial community data
  - Upload for reads and assemblies
  - Extensive QC
- Automated processing and analysis of
  - ITS / 16s / 18s
  - shotgun metagenome
  - meta transcriptome
- Comparison
  - Functional
  - Taxonomic
  - Using precomputed profiles
- Subset retrieval
  - All reads for *E.coli*
  - All reads for Lysine Biosynthesis
  - All reads hitting dnaA



# of metagenomes	72,795
# base pairs	22.4 Tbp
# of sequences	206.07 billion
# of public metagenomes	12,243

# What problems do we solve?

- Analysis of single shotgun metagenomics
  - With “large data”
    - Don't tell any physicist I said that
  - Comparison of MANY data sets
    - **Note:** analyzed data is 10x the size of input data
  - **We need to be very efficient with resources**
    - We do not use big iron!
    - Had to improve pipeline 750x over past 18months
- ➔ Triggered infrastructure research projects  
(SHOCK and AWE)

# Metagenome Upload (ftp, http, gridFTP)

↑ Data Submission

3. select sequence file(s)

4. choose pipeline options

## SELECTED PIPELINE OPTIONS

**assembled**  Select this option if your input sequence file(s) contain assembled data and include the coverage information within each sequence header as described [here](#).

**dereplication**  Remove artificial replicate sequences produced by sequencing artifacts [Gomez-Alvarez, et al, The ISME Journal \(2009\)](#).

**screening**

Remove any host specific species sequences (e.g. plant, human or mouse) using DNA level matching with bowtie [Langmead et al., Genome Biol. 2009, Vol 10, issue 3](#)

**dynamic trimming**  Remove low quality sequences using a modified DynamicTrim [Cox et al., \(BMC Bioinformatics, 2011, Vol. 11, 485\)](#).

(fastq only)

Specify the lowest phred score that will be counted as a high-quality base.

Sequences will be trimmed to contain at most this many bases below the above-specified quality.

**length filtering**  Filter based on sequence length when no quality score information is available.

(fasta only)

Specify the multiplier of standard deviation for length cutoff.

**ambiguous base filtering**  Filter based on sequence ambiguity base (non-ACGT) count when no quality score information is available.

(fasta only)

Specify the maximum allowed number of ambiguous basepairs.

select

**Warning: Comparison of datasets processed with different pipeline options may not be valid.**

# Metagenome Overview

## Metagenome Overview

**MG-RAST ID** 4447970.3 Download Analyze Search

**Metagenome Name** CA\_05\_4.6  
**PI** Alex Mira  
**Organization** CSISP  
**Visibility** Public  
**Static Link** <http://metagenomics.anl.gov/linkin.cgi?metagenome=4447970.3>

**NCBI Project ID** -  
**GOLD ID** -  
**PubMed ID** 21716308

### METAGENOME SUMMARY

Dataset CA\_05\_4.6 was uploaded on 05/05/2010 and contains 70,503 sequences totaling 27,669,924 basepairs with an average length of 392 bps. The piechart below breaks down the uploaded sequences into 5 distinct categories.

0 sequences (0.0%) failed to pass the QC pipeline. Of the sequences that passed QC, 1,569 sequences (2.2%) contain ribosomal RNA genes. Of the remainder, 48,591 sequences (68.9%) contain predicted proteins with known functions and 13,998 sequences (19.9%) contain predicted proteins with unknown function. 6,345 sequences (9.0%) have no rRNA genes or predicted proteins.

The analysis results shown on this page are computed by MG-RAST. Please note that authors may upload data that they have published their own analysis for, in such cases comparison within the MG-RAST framework can not be done.

- DOWNLOAD data and annotations
- ANALYZE annotations in detail.
- SEARCH through annotations.

#### Sequence Breakdown

### TABLE OF CONTENTS

- Work with Metagenome Data
  - [Download](#)
  - [Analyze](#)
  - [Search](#)
- Overview of Metagenome
  - [Summary](#)
  - [Project Information](#)
  - [GSC MixS Info](#)
  - [Publication Abstracts](#)
- Metagenome QC
  - [DRISEE](#)
  - [Kmer Profile](#)
  - [Nucleotide Histogram](#)
- Organism Breakdown
  - [Taxonomic Distribution](#)
  - [Rank Abundance Plot](#)
  - [Rarefaction Curve](#)
  - [Alpha Diversity](#)

# Comparison tools

## Metagenome Analysis

### 1 Data Type

ORGANISM ABUNDANCE

- Representative Hit Classification
- Best Hit Classification
- Lowest Common Ancestor

FUNCTIONAL ABUNDANCE

- » Hierarchical Classification
- All Annotations

OTHER

- Recruitment Plot

### 2 Data Selection

Metagenomes

4447970.3, 4447971.3, 4440284.3, 4440285.3, 4440286.3, 4440055.3, 4440056.3, 4440059.3, 4440066.3, 4440062.3, 4440063.3

Annotation Sources

Max. e-Value Cutoff: 1e-5

Min. % Identity Cutoff: 60 %

Min. Alignment Length Cutoff: 15

Workbench  use features from workbench

### 3 Data Visualization

barchart  tree  table  heatmap  PCoA

**KEGG Mapper**

### Workbench (0 Features) Getting Started

To create a visualization, first select an analysis view from the **Analysis Views** box. The default is 'Organism Classification'. Then choose the data and cutoffs you wish to use in the **Data Selection** box. Depending on the type of data, you might have a set of possible visualizations. Pick one of them and click the **generate** button.

You will see the generated visualizations created in separate tabs in this tab-view. In addition to the visualization

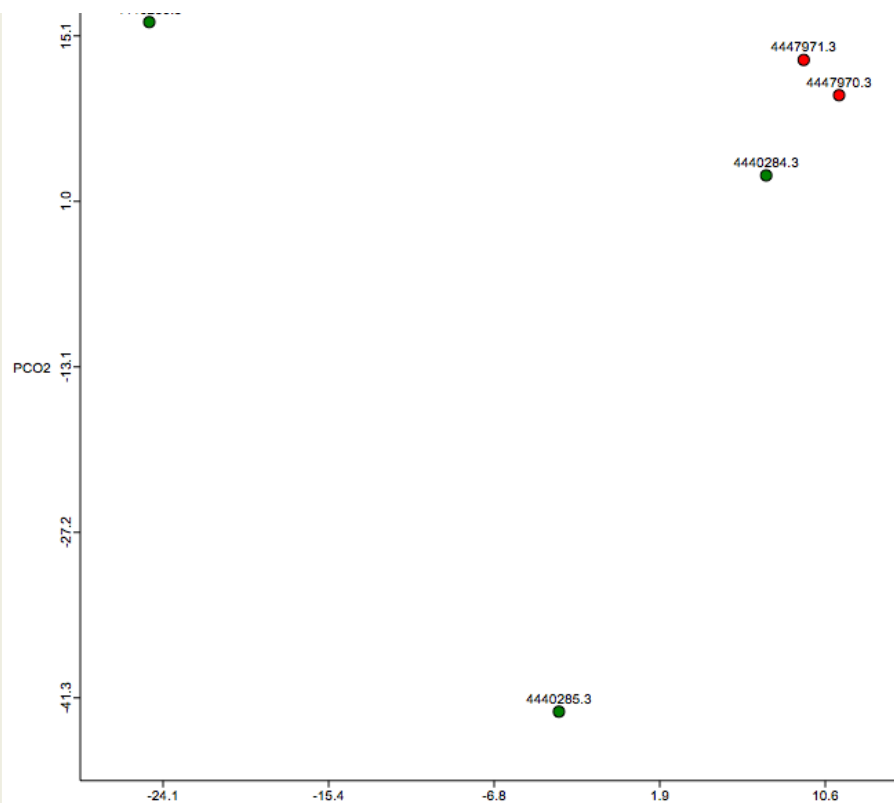


This data was calculated for metagenomes 4447970.3, 4447971.3, 4440284.3, 4440285.3 and 4440286.3. The data was compared to Subsystems using a maximum e-value of 1e-5, a minimum identity of 60 %, and a minimum alignment length of 15 measured in aa for protein and bp for RNA databases. The data has been normalized to values between 0 and 1. If you would like to view raw values, redraw using the form below.

redraw using  values and  distance

The image is currently dynamic. To be able to right-click/save the image, please click the static button

**PCoA grouping control**



Component	%	X-axis	Y-axis
PCO1	0.61620	<input checked="" type="radio"/>	<input type="radio"/>
PCO2	0.23509	<input type="radio"/>	<input checked="" type="radio"/>
PCO3	0.13954	<input type="radio"/>	<input type="radio"/>
PCO4	0.00918	<input type="radio"/>	<input type="radio"/>
PCO5	0.00000	<input type="radio"/>	<input type="radio"/>

group	name	save as collection
group 1	<input type="text" value="human"/>	<input type="button" value="save"/>
group 2	<input type="text" value="chicken"/>	<input type="button" value="save"/>
group 3	<input type="text"/>	<input type="button" value="save"/>
group 4	<input type="text"/>	<input type="button" value="save"/>
group 5	<input type="text"/>	<input type="button" value="save"/>
group 6	<input type="text"/>	<input type="button" value="save"/>
group 7	<input type="text"/>	<input type="button" value="save"/>
group 8	<input type="text"/>	<input type="button" value="save"/>
group 9	<input type="text"/>	<input type="button" value="save"/>
group 10	<input type="text"/>	<input type="button" value="save"/>

# MG-RAST Notebooks / IPython integration



MGNB - The MG-RAST Notebook

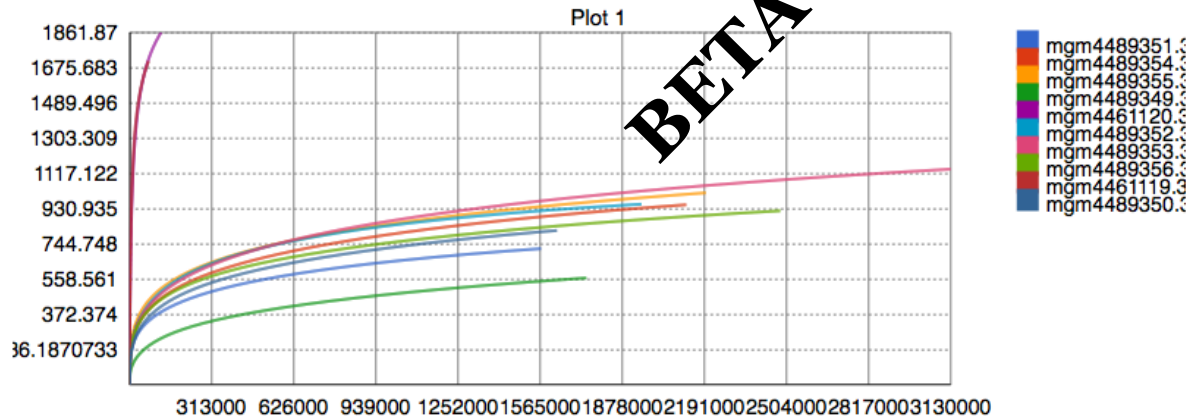
Notebook 1 ✕ +

Analysis Builder

File Edit Insert Cell Kernel Help

```
In [17]: # my data subselection
selected_ids = [ 'mgm4461119.3', 'mgm4461120.3', 'mgm4489349.3', 'mgm4489350.3', 'mgm4489351.3', 'mgm4489352.3', 'mgm4489353.3', 'mgm4489354.3', 'mgm4489355.3', 'mgm4489356.3' ]
rare_args_1 = sample_set['statistics'].plot_rarefaction(mgids=selected_ids, arg_list=True)
```

```
In [18]: rare_args_1.update({'show_legend': True, 'legend_position': 'right', 'connected': True, 'show_dots': False, 'title': 'Plot 1'})
Ipy.RETINA.plot(**rare_args_1)
```



BETA TEST

# Data access / Downloads

## Project Overview

**THE ORAL METAGENOME IN HEALTH AND DISEASE (ID 128)**

Visibility: Public  
Static Link: <http://metagenomics.anl.gov/linkin.cgi?project=128>

Share Project | Add Jobs | Edit Project Data | Upload Info | Upload MetaData | Export MetaData

MG-RAST ID	Metagenome Name	bp Count	Sequence Count	Biome	Feature	Material	Location	Country	Coordinates	Sequence Type	Sequence Method	Download
		<	<	huma	human	human				WGS	454	
4447943.3	CA_04P	142,374,233	339,503	human-associated habitat	human-associated habitat	human-associated habitat	Valencia	Spain	39.481448, 0.353066	WGS	454	
4447192.3	NOCA_01P	77,538,485	204,218	human-associated habitat	human-associated habitat	human-associated habitat	Valencia	Spain	39.481448, 0.353066	WGS	454	
4447103.3	CA1_01P	203,711,161	464,594	human-associated habitat	human-associated habitat	human-associated habitat	Valencia	Spain	39.481448, 0.353066	WGS	454	
4447102.3	NOCA_03P	100,125,112	244,881	human-associated habitat	human-associated habitat	human-associated habitat	Valencia	Spain	39.481448, 0.353066	WGS	454	
4447101.3	CA1_02P	129,851,692	295,072	human-associated habitat	human-associated habitat	human-associated habitat	Valencia	Spain	39.481448, 0.353066	WGS	454	
4447971.3	CA_06_1.6	37,519,874	97,722	human-associated habitat	human-associated habitat	human-associated habitat	Valencia	Spain	39.481448, 0.353066	WGS	454	
4447970.3	CA_05_4.6	27,669,924	70,503	human-associated habitat	human-associated habitat	human-associated habitat	Valencia	Spain	39.481448, 0.353066	WGS	454	
4447903.3	CA_06P	123,266,763	306,740	human-associated habitat	human-associated habitat	human-associated habitat	Valencia	Spain	39.481448, 0.353066	WGS	454	

# Download // data access

- Data for all but one visual can be downloaded as spreadsheet
- arbitrary subset can downloaded
  - In standard formats
- All data is available
  - For public projects
- Data is typically private
  - User decides to publish
- **Pre-publication** sharing via email-token

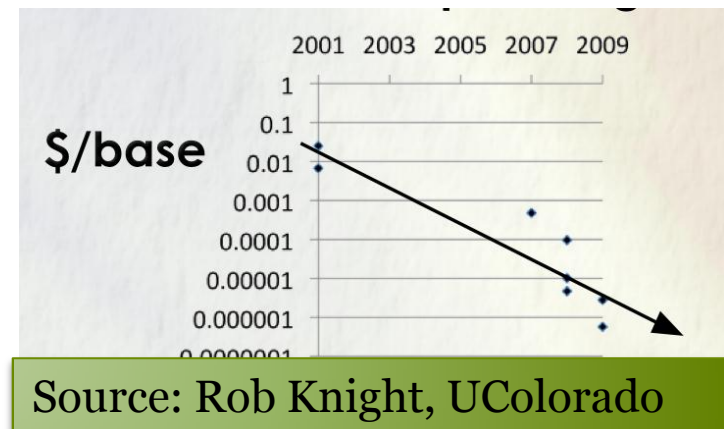
# Reducing Terabytes to manageable objects

Terabases → 100,000 protein functions with abundance → 4 profiles (COG, SEED, KEGG, NOG)  
→ 900,000 organisms → 1 profiles (NCBI taxonomy)  
→ ability to go back to sequences

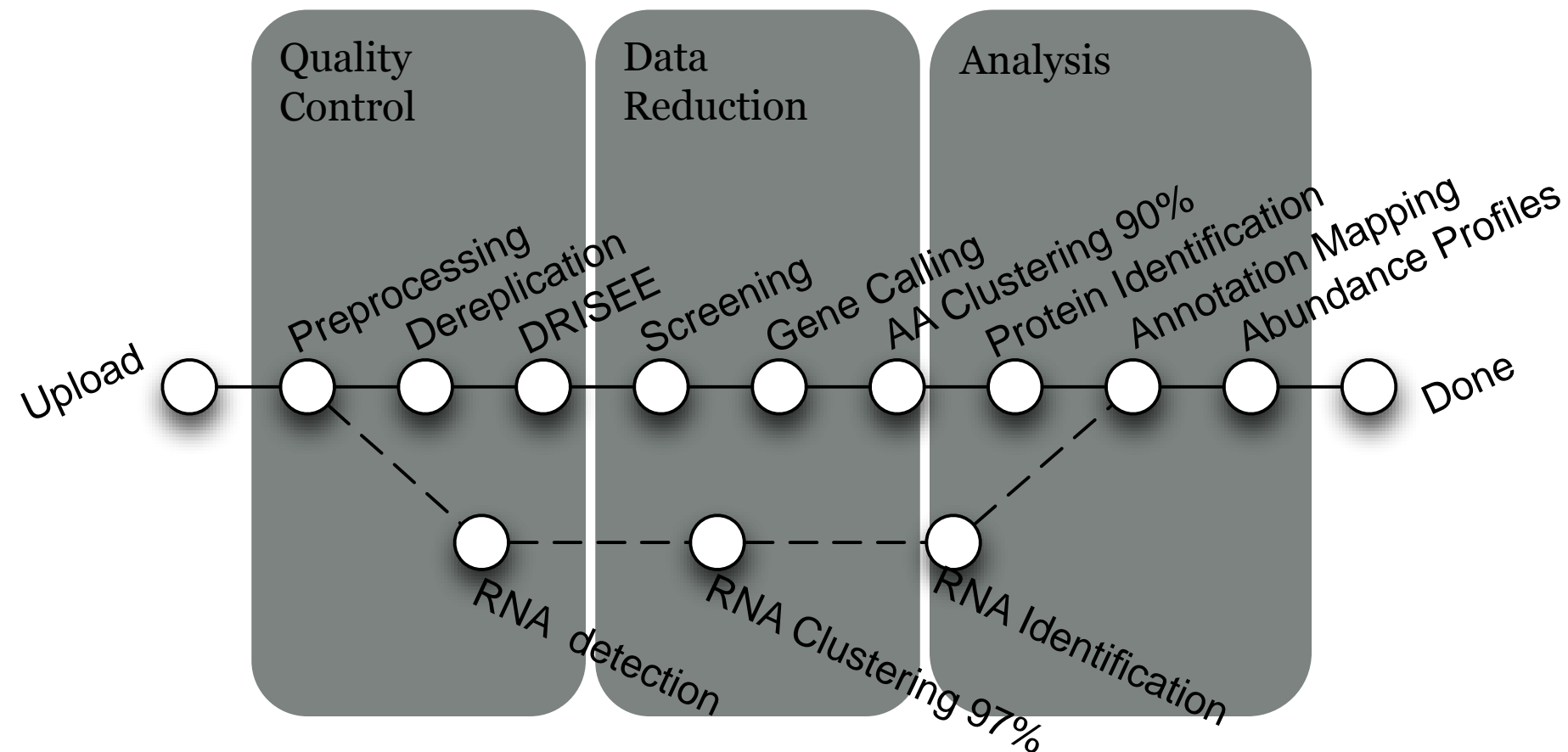
- Data volume growing fast:
  - 2003: C. Venter's GOS with **600MBp** (or 0.6GBp)
  - 2011: HMP with **6TBp** (or 6,000GBp)
- Data is different: shorter reads, but many reads, noisy
- Major cost is in bioinformatics (10x the sequencing cost today)



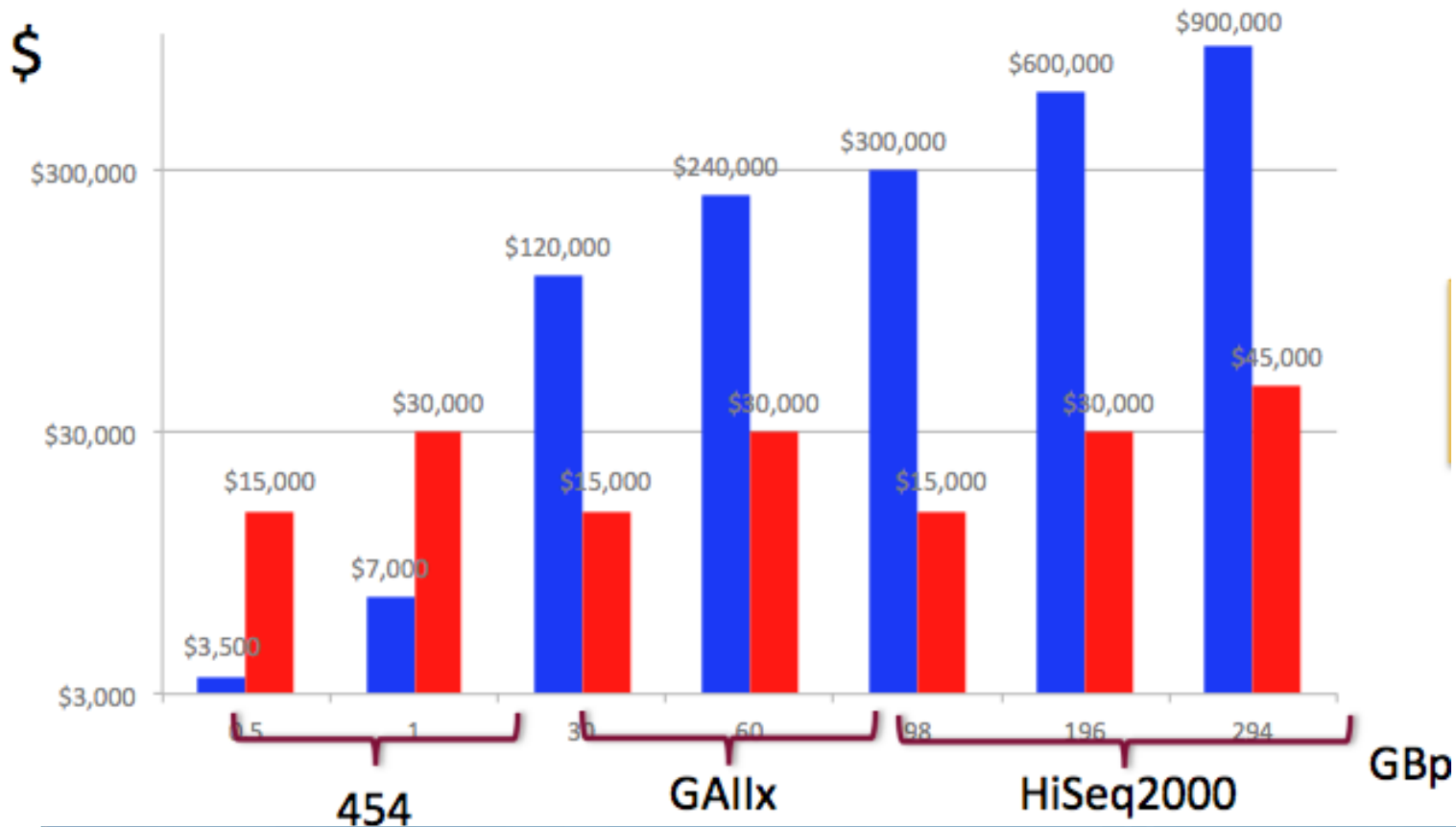
Driving force



# The MG-RAST v3.x pipeline



# Analysis cost are dominating



**Computing cost (blue) dominate sequencing (red)**

- Cost on Amazon EC2 Cloud, September 2009

- Pure run-time for BLASTX, no storage or data transfer

# No one size fits all

- Every experiment is different
- Every sample is different
- ➔ MG-RAST does not provide one size fits all
  - No single download for all purposes.
- We allow users to change parameters for analysis at view time
  - Using intelligent data products to construct annotated reads for specific parameters
- ➔ Web interfaces ties it all together



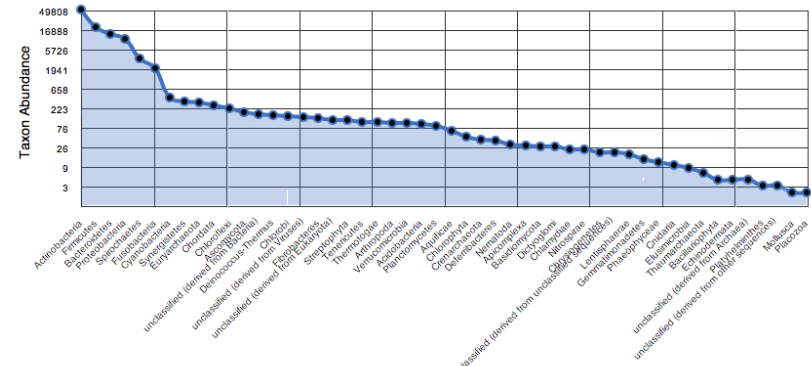
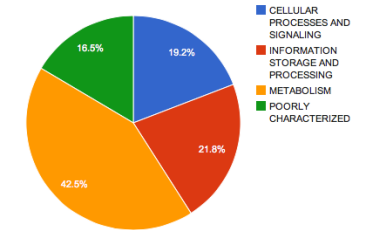
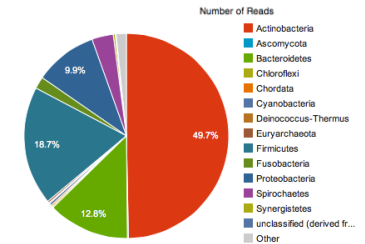
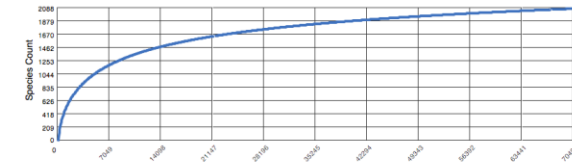
# Common views

Rarefaction

Taxonomic breakdown

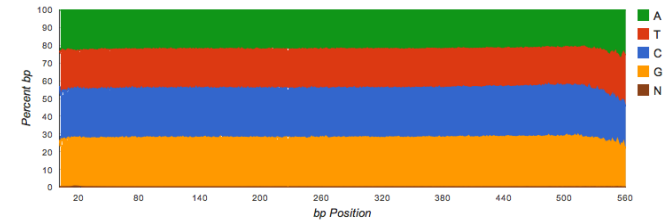
Functional breakdown

Rank abundance

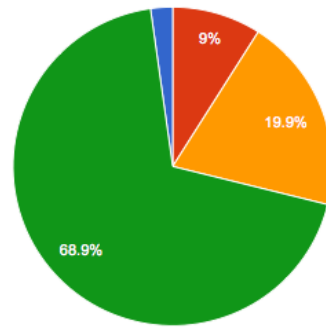


# More views

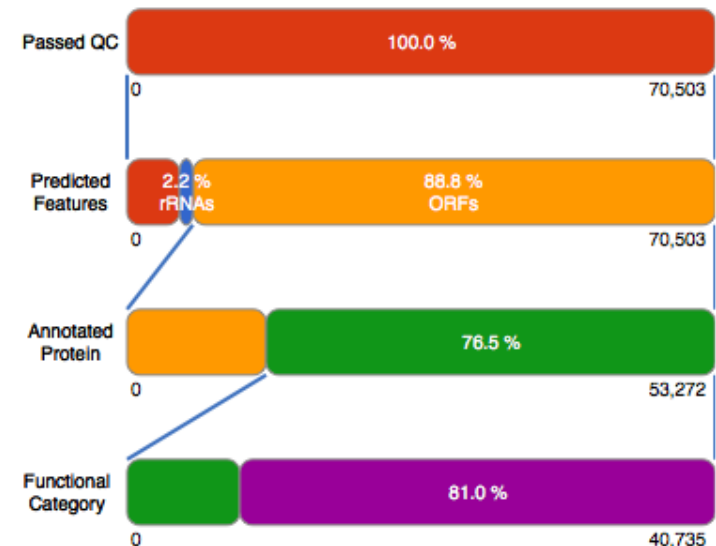
- Nucleotide histogram



- QC

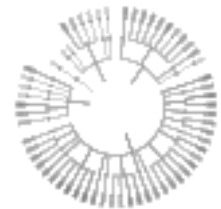
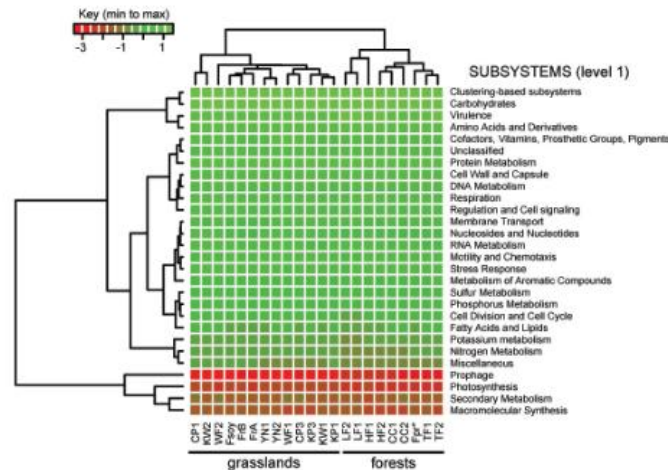
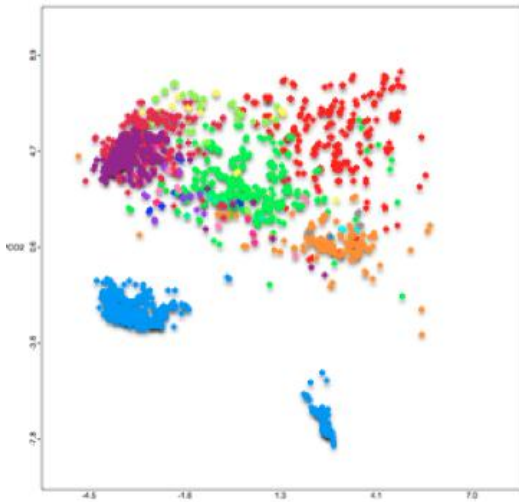
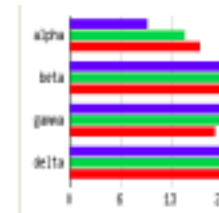
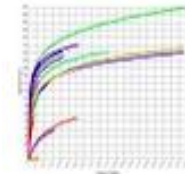


- Processing



# Comparative tools

- Comparison of many data sets
- ➔ **Normalization** allows comparison
  - Many name spaces (GenBank, SEED, KEGG, GO, COG, eggNOG, UniProt supported)
  - Parameters can be varied at query time



# Lessons learned...

- Old style:
  - “Lets sequence as much as we can afford”
  - “Metagenomics is like genomics”
- Today:
  - Often 16s amplicon study first
  - replicates (biological and technical)
    - “design for statistics”
    - “replicate or lie” (Jim Prosser)
  - metadata
    - Genomics Standards Consortium provides tools
  - Provide good QC
    - Identify signal vs. noise ratio
    - Throw away bad data when needed (!)
  - Identify appropriate analysis workflow
  - Perform assembly?

**→ Design for statistics**

**→ Metadata  
(r)evolution**

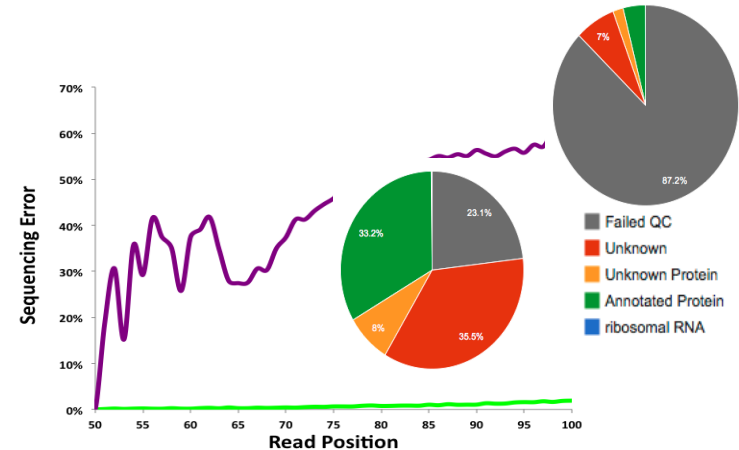
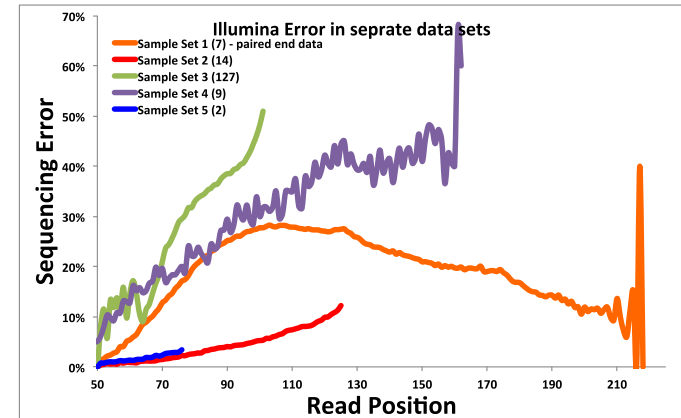
**→ Data hygiene**

**→ Tool chain matters**

# DRISEE - objective QC for NGS data

- Approach is simple
  - Develop synthetic reference
- DRISEE:
  - using ADRs to find noise
  - Correlates well with our ability to analyze data

➔ Numerous quality issues  
➔ Not all NGS data is alike  
➔ Even from the same vendor

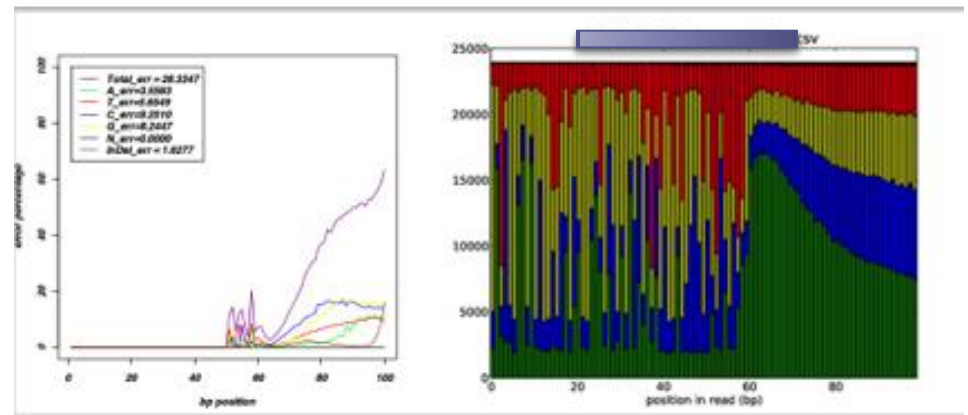
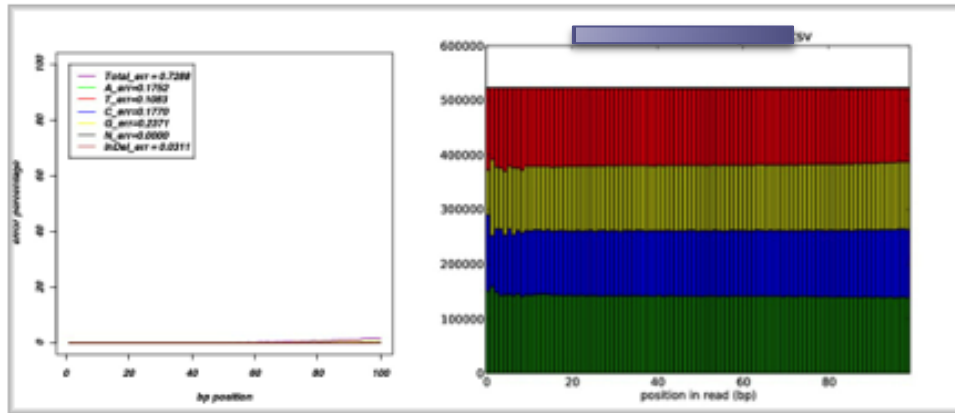


Find prefix identical subsets



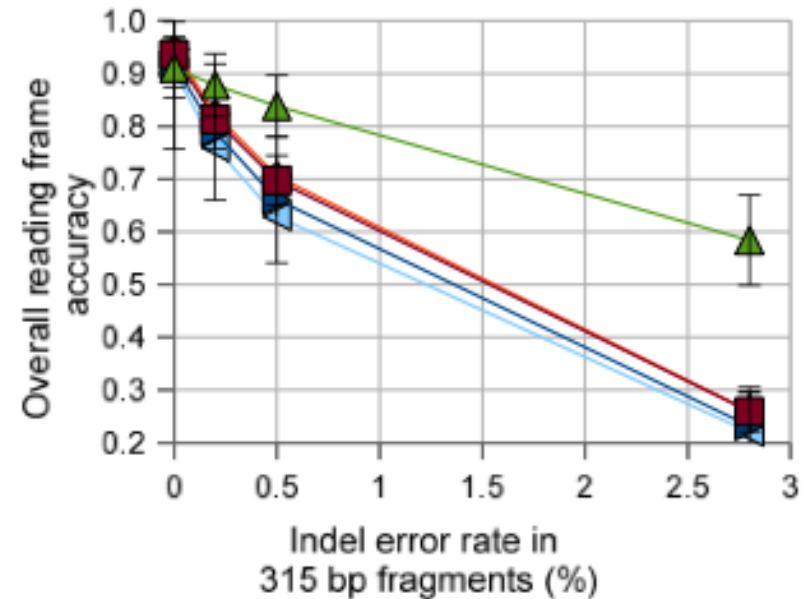
➔ Discrepancies are error

# Extreme cases of good and bad data quality



# Tool chain variation

- Question:
  - What happens if I vary the tool chain?
- **Existing approaches** rely on:
  - Compare results of different studies (ie multiple pipelines)
- Here we study 5 different popular gene finding tools for metagenome on **simulated data**
- **Effects are dramatic**
  - Accuracy goes drops dramatically with moderate error
- Comparison of data requires identical tool chain



From: Trimble et al, BMC Bioinformatics, 2012



# Analysis tool chain



- (small) changes in the analysis tool change will have dramatic impact on results
- Comparing the results of two independently analyzed studies is next to impossible
  - We need to “normalize” analysis
  - MG-RAST currently has 72,000 normalized metagenomes for comparison
  - Over ~11k are public

**MG-RAST**  
metagenomics analysis server



# Programmatic interface (API)

- Part of the KBase API
  - Contributes some unique features
- Uses standards whenever possible
  - Metadata (GSC MlxS)
  - Abundance profiles (BIOM)
- Enables extension by third parties
- Data download and subselection

## API Details

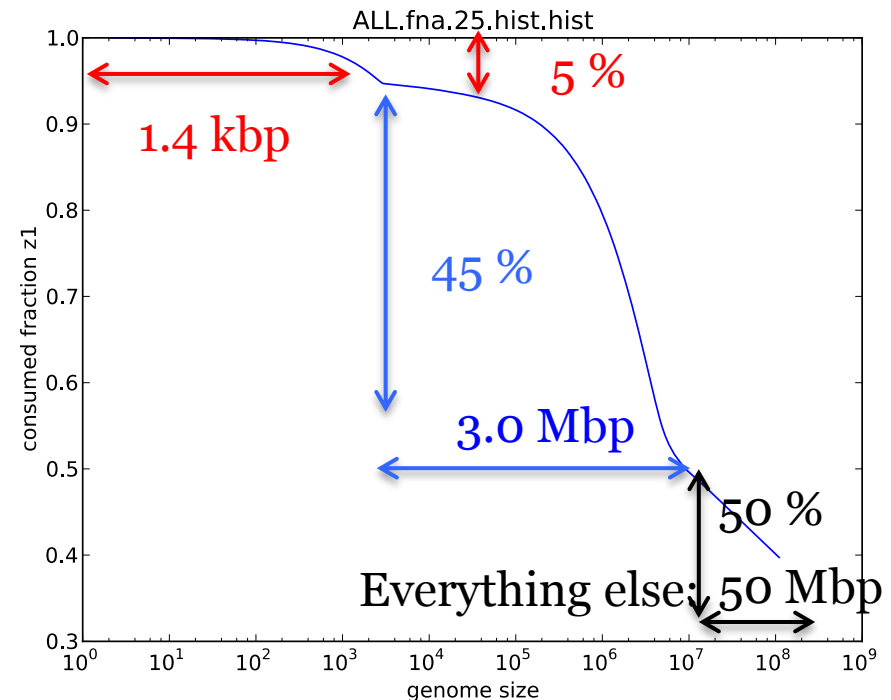
- Added 1600 function calls
- Grouped into 14 higher level objects
- Currently access to 11,000 public metagenomes
- Subsetting by function and taxonomy
- Supports KEGG, SEED, UniProt, GenBank, IMG, COG, eggNOG, RDP, SILVA similarities

**Download (ftp) and RESTful**

**Status:** Public beta

# Predicting replicons before assembly

- Using **K-mer spectra** to predict (pan-) genome size
  - K-mer= unique word, easily computed
- In addition to alpha diversity
  - 300 OTU data set
- Using k-mer size 25
- **Red** and **blue** replicons were missing in assembly
  - Allows adjustment of parameters



Thank you very much for you attention

